САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

И МНОГОПРОЦЕССОРНЫХ СИСТЕМ

Христенко Евгений Александрович

Выпускная квалификационная работа бакалавра

" Сравнение эффективности методов многомерной визуализации"

Направление 010400

Прикладная математика и информатика

Научный руководитель,
кандидат тех. наук,
доцент
Мареев В.В.

Санкт-Петербург

2016

Содержание

Введение

Совершенствование методов, используемых в области социологических, политических, экономических исследований непрерывно ведет к усложнению и увеличению количества полученных данных. В связи с этим возникает ряд проблем при анализе таких данных. Например, возникают трудности с однозначным определением распределения многомерных данных, которые могут быть распределены не нормально, а, допустим, фрактально. Данная работа посвящена решению одной из существующих проблем, а именно задаче сечения исходных данных большой размерности.

С точки зрения математической статистики, такие данные характеризуются большим количеством параметров. Вследствие чего, анализ требует большого количества вычислений и, следовательно, использования современных информационных технологий и специализированного программного обеспечения. Показательным примером «сложных инструментов» анализа являются многомерные методы. Многомерные методы – наиболее трудоемкие и ресурсозатратные (с точки зрения расчетов) методы в математической статистике. Однако, не редки случаи, когда исследователь не располагает необходимой материальной базой. Ввиду отсутствия значительных средств и доступа к большим вычислительным мощностям, ученый вынужден анализировать двухмерные сечения исходных данных как наиболее простые объекты исследования, т. е. попытаться установить зависимость конкретной переменной от одного из возможных параметров.

Целью работы является разработка некоторого алгоритма для исследования взаимосвязи интересующего нас явления с множеством всех параметров.

Кроме того, в работе рассматривается вопрос визуализации полученных результатов для возможности дальнейшей интерпретации.

Для решения этой задачи ресурсным центром «Вычислительный центр СПбГУ» была предоставлена исследовательская платформа в виде виртуального вычислительного сервера на одном из вычислительных кластеров центра, а также статистический пакет программ IBM SPSS Statistics версии 21 в качестве основного средства анализа [1].

В качестве исходных данных были взяты данные, предоставленные Европейским Социальным Исследованием (the ESS) [2]. Европейское социологическое исследование – это двухлетняя работа, цель которой слежение за изменениями в институтах, предпочтениях, уверениях и поведенческих моделях людей в Европе. Начатое в 2002 году исследование проводилось каждые два года во многих европейских странах. Были взяты результаты для Российской Федерации в 2012 году, полученные в ходе шестой волны исследования [2]. База данных ESS хорошо подходит для исследования в рамках поставленной задачи, так как содержит большое количество переменных. В изначальном варианте в ней насчитывается 626 переменных. После исключения пустых полей, отвечающих за граждан других стран, а также полей, содержащих незначительное количество информации, база содержит 241 переменных. Именно этот вариант был принят в качестве исходных данных.

Постановка задачи

В соответствие с целью работы можно выделить две большие подзадачи:

1. Исследование данных при помощи статистических методов;
2. Визуализация полученных результатов;

В рамках первого пункта были сформулированы следующие задачи:

- выбрать исследуемую переменную;
- при помощи корреляционного анализа отобрать параметры наиболее значительные для исследуемого параметра;
- построить многомерную регрессионную модель для всех параметров и множество одномерных регрессионных моделей для каждой пары «зависимая переменная — независимая переменная» и сравнить получившиеся результаты;
- при помощи факторного анализа снизить размерность данных для дальнейшей визуализации.

Решение вышеперечисленных задач описано в главе 1.

Глава 2 посвящена вопросам визуализации результатов полученных в главе 1.

Обзор литературы

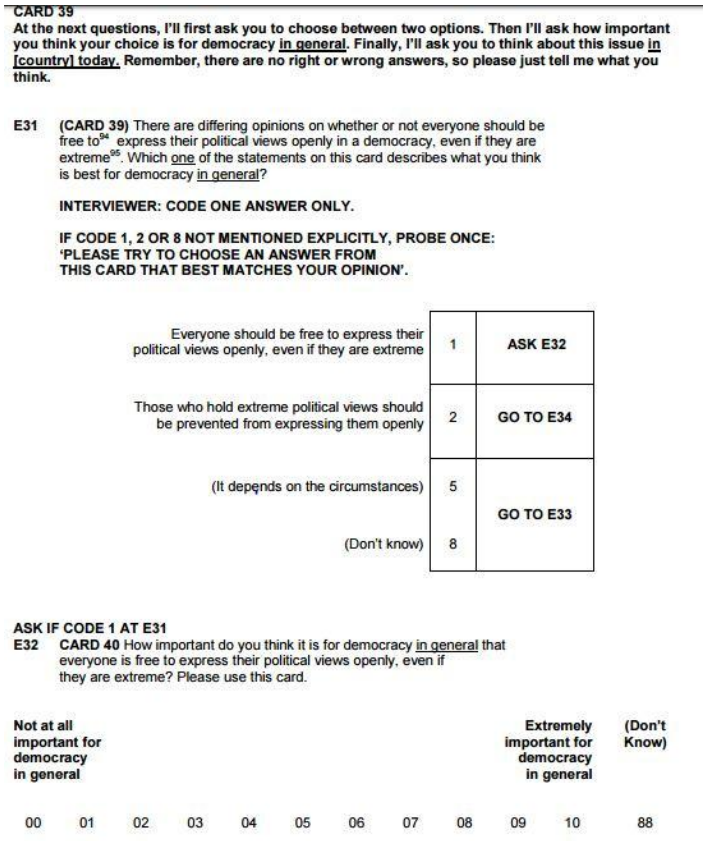
Так как исследования производились с использованием среды SPSS Statistics, была изучена литература, содержащая информацию о принципах работы в данной системе и возможностях программы [3]. Из книги [4] были почерпнуты общие представления о многомерном статистическом анализе. Подробную информацию обо всех многомерных статистических методах, использованных в работе можно получить из книги [5]. Формулы для расчета коэффициентов корреляции и проверки гипотезы о значимости таких коэффициентов были взяты из книги [6]. Порядковая регрессия, рассмотренная в параграфе 1.4, была построена на основе статьи [7], а факторный анализ из параграфа 1.5 – на основе статьи [8] и книги [9].

Для решения задачи визуализации была прочитана книга [10], которая дает общее представление о визуализации многомерных данных. Кроме того, ценным источником информации по данной теме является статья [11].

Глава 1. Статистический анализ данных ESS

1.1. Выбор исследуемого параметра

ESS – это всестороннее исследование социального уровня и благосостояния отдельно взятой страны. Опросы содержат большое количество различных вопросов, затрагивающих все области социальной жизни человека. Как следствие, имеется значительный выбор для исследователя в предмете изучения. На рисунке 1 приведен фрагмент опроса, иллюстрирующий сложность, детальность и проработанность методов, используемых в ESS.



CARD 39
At the next questions, I'll first ask you to choose between two options. Then I'll ask how important you think your choice is for democracy in general. Finally, I'll ask you to think about this issue in [country] today. Remember, there are no right or wrong answers, so please just tell me what you think.

E31 **(CARD 39)** There are differing opinions on whether or not everyone should be free to[94] express their political views openly in a democracy, even if they are extreme[95]. Which one of the statements on this card describes what you think is best for democracy in general?

INTERVIEWER: CODE ONE ANSWER ONLY.

IF CODE 1, 2 OR 8 NOT MENTIONED EXPLICITLY, PROBE ONCE: 'PLEASE TRY TO CHOOSE AN ANSWER FROM THIS CARD THAT BEST MATCHES YOUR OPINION'.

Everyone should be free to express their political views openly, even if they are extreme	1	**ASK E32**
Those who hold extreme political views should be prevented from expressing them openly	2	**GO TO E34**
(It depends on the circumstances)	5	**GO TO E33**
(Don't know)	8	

ASK IF CODE 1 AT E31
E32 **CARD 40** How important do you think it is for democracy in general that everyone is free to express their political views openly, even if they are extreme? Please use this card.

Not at all important for democracy in general										**Extremely important for democracy in general**	**(Don't Know)**
00	01	02	03	04	05	06	07	08	09	10	88

Рисунок 1. Фрагмент опроса ESS

В данной работе в качестве объекта исследования был выбран параметр «Насколько Вы удовлетворены своей жизнью в целом», так как именно этот параметр наиболее просто и доступно отражает уровень развития страны и благосостояния граждан, а также способен агрегировать в себе прочие

показатели. Он принимает значения от 0 (полностью недовольны) до 10 (абсолютно удовлетворены).

Ниже на рисунке 2 приведена гистограмма избранного параметра для предварительного представления об исследуемом объекте, а также отображена нормальная кривая для поверхностной оценки нормальности распределения.

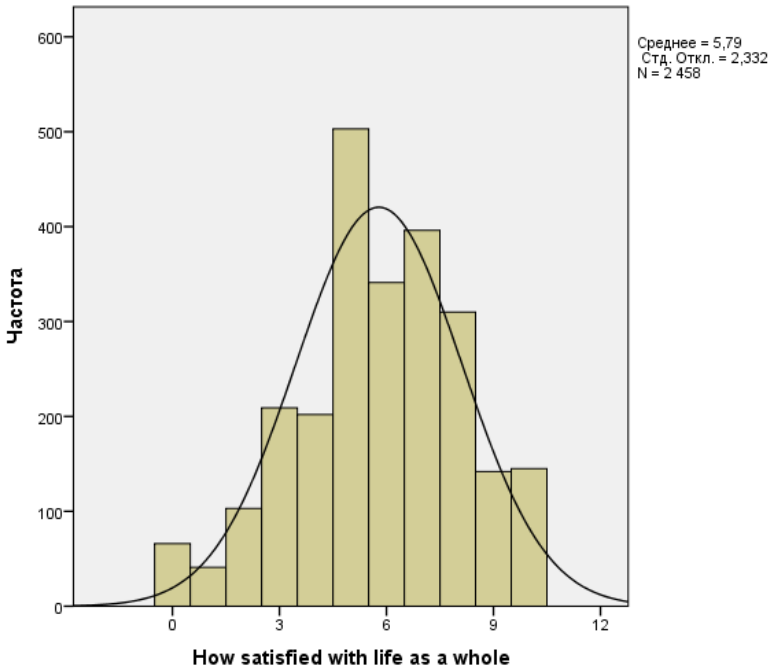


Рисунок 2. Гистограмма переменной «Насколько вы удовлетворены своей жизнью в целом»

1.2. Выявление зависимости между параметрами

Для решения задачи о выявлении зависимости между зонами были вычислены выборочные коэффициенты корреляции для всех пар «зависимая переменная − независимая переменная» по формуле [6]:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sigma_x\sigma_y},$$

Где $\overline{x}$ и $\overline{y}$ – выборочные средние, а σ_x^2, σ_y^2 – выборочные дисперсии, вычисленные по первой и второй выборкам соответственно.

При уровне значимости α =0,05 необходимо проверить гипотезу H_0 о значимости коэффициентов корреляции.

В качестве нулевой принимаем гипотезу о значимости коэффициента корреляции. Для проверки гипотезы вычислим значения статистик по формуле [6]:

$$t_r = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}.$$

1. Если $|t_r| \geq t_{n-2}$, то гипотеза H_0 принимается;
2. Если $|t_r| < t_{n-2}$, то гипотеза H_0 следует отклонить.

Значение t_{n-2} определяется по таблице распределения Стьюдента при n - 2 степенях свободы.

Кроме того, полагаем, что существенными для нас будут параметры, коэффициент корреляции которых превосходит 0,3. Из 241 признака такому критерию будут удовлетворять только восемь параметров, перечисленных в таблице 1(также указаны соответствующие им сокращения).

Название	**Сокращение**	**Принимаемые значения**
«Насколько Вы удовлетворены состоянием экономики»	eco	От 0 до 10
«Насколько Вы счастливы»	hap	От 0 до 10
«Как Вы оцениваете свое здоровье»	hea	От 1 до 9
«Чувствуете, что близкие Вас ценят»	clo	От 0 до 10
«Как часто заинтересованы тем, что Вы делаете»	int	От 0 до 10
«Обращаете внимание и оцениваете Ваше окружение»	sur	От 0 до 10
«Есть чувство направленности в Вашей жизни»	dir	От 0 до 10
«Насколько Вы удовлетворены своей работой»	job	От 0 до 10

Таблица 1. Отобранные переменные

В таблице 2 содержатся соответствующие коэффициенты корреляции. Из нее видно, что все независимые переменные слабо коррелируют с зависимой переменной, за исключением второго пункта, что согласуется с действительностью, поскольку параметр, отражающий счастье респондента, по своей сути во многом похож на исследуемый нами признак.

eco	hap	hea	clo	int	sur	dir	job
0,354	0,696	0,307	0,350	0,333	0,306	0,356	0,335

Таблица 2. Коэффициенты корреляции

1.3. Построение регрессионных моделей

Так как зависимая переменная имеет порядковую меру, классическая модель линейной регрессии становится неприменима. Был применен метод порядковой регрессии, реализация которого присутствует в SPSS [7].

Порядковая регрессия – это расширение обобщенной линейной модели регрессии, в которой зависимая переменная измеряется в порядковой шкале. Независимые переменные в модели порядковой регрессии могут быть категориальными или количественными. Категориальные независимые переменные называют факторами. А количественные независимые переменные – ковариатами.

В модели порядковой регрессии для каждой категории зависимой порядковой переменной (за исключением последней) строится уравнение регрессии, прогнозирующее накопленную вероятность принадлежности объекта наблюдения к данной категории.

В качестве связывающей функции был использован сопряженный двойной логарифм, так как в соответствие с рисунком 2 более вероятны высокие значения зависимой переменной.

Была построена многомерная порядковая регрессионная модель на основе всех выделенных параметров, а также множество одномерных порядковых регрессионных моделей для каждой из 8 независимых переменных по отдельности. В приложении 1 содержатся результаты построения.

Многомер.	eco	hap	hea	clo	int	sur	dir	job
0,328	0,210	0,308	0,202	0,206	0,207	0,210	0,206	0,207

Таблица 3. Точность регрессионных моделей

В таблице 3 показана точность получившихся моделей. Значения хорошо иллюстрируют преимущество многомерного подхода над одномерным подходом. Многомерная модель имеет существенно большую точность, нежели одномерные модели. Только модель, основанная на сильно коррелирующем параметре hap, имеет ожидаемо большую степень точности, сопоставимую с точностью многомерной модели, однако все же уступает ей. Поэтому при исследовании данных, содержащих большое количество переменных, предпочтительно использовать многомерные методы. Анализ же одномерных моделей может привести исследователя к ошибочному результату.

1.4. Снижение размерности данных

Для снижения размерности исходных данных воспользуемся возможностями факторного анализа, а именно методом главных компонент [8], [9]. В качестве метода вращения был выбран метод «варимакс». На рисунке 3 представлены результаты анализа, построенного в среде SPSS Statistics.

Значение напротив переменной называется факторной нагрузкой. Эта величина означает корреляцию между исходной переменной и компонентом

(фактором). В соответствие с наибольшим абсолютным значением нагрузки переменные разделяются на 3 группы соответственно каждому фактору:

	Компонент		
	1	2	3
How satisfied with life as a whole	,367	,558	,449
How satisfied with present state of economy in country	,023	,074	,910
How happy are you	,377	,619	,315
Subjective general health	-,011	-,845	,067
Feel appreciated by people you are close to	,709	,268	-,132
Interested in what you are doing, how much of the time	,769	,092	-,002
Take notice of and appreciate your surroundings	,675	-,054	,284
Have a sense of direction in your life	,703	,142	,083
How satisfied with job	,511	,241	,177

Рисунок 3. Повернутая матрица компонентов

1. «Чувствуете, что близкие Вас ценят», «Как часто заинтересованы тем, что Вы делаете», «Обращаете внимание и оцениваете Ваше окружение», «Есть чувство направленности в Вашей жизни», «Насколько Вы удовлетворены своей работой»;
2. «Насколько Вы удовлетворены своей жизнью в целом», «Насколько Вы счастливы», «Как вы оцениваете свое здоровье»;
3. «Насколько Вы удовлетворены состоянием экономики».

Первый компонент собрал в себе менее значительные, частные субъективные положения. Во второй компонент входят более значительные, общие субъективные вопросы. Третий компонент можно интерпретировать

как оценки респондентом внешних условий, не относящихся к жизни конкретного индивида.

Необходимо убедится в справедливости проведенного разбиения. Факторные переменные принимают значения от -3 до 3. Перейдём к рассмотрению третьего наблюдения, значение факторов которого соответственно равно:

-0,60501 2,03241 0,96567

Как следствие, ожидаются достаточно высокие значения для параметров второго компонента (за исключением параметра «Как Вы оцениваете свое здоровье», который, напротив, должен иметь низкое значение, так как входит в компонент с отрицательной нагрузкой) и значения немного ниже и немного выше среднего для первого и третьего компонентов соответственно. В справедливости такой оценки можно убедится ознакомившись с данными, представленными в таблице 4.

Название	**Фактор**	**Принимаемое значение**
«Чувствуете, что близкие Вас ценят»	1	5
«Как часто заинтересованы тем, что Вы делаете»		7
«Обращаете внимание и оцениваете Ваше окружение»		4
«Есть чувство направленности в Вашей жизни»		3
«Насколько Вы удовлетворены своей работой»		3
«Насколько Вы счастливы»	2	10
«Как Вы оцениваете свое здоровье»		1
«Насколько Вы удовлетворены своей жизнью в целом»		5
«Насколько Вы удовлетворены состоянием экономики»	3	7

Таблица 4. Значения переменных третьего наблюдения

1.5. Выводы по первой главе

В этой главе был проведен статистический анализ базы данных ESS. В ходе работы были получены следующие результаты:

1. На удовлетворенность жизнью человека, согласно Европейскому Социальному Исследованию в большей степени оказывают влияние восемь выше перечисленных параметров;
2. Многомерная регрессионная модель, построенная на основе таких параметров, способна по ответам респондента предсказать его удовлетворенность жизнью с точностью 0,328;
3. При помощи метода главных компонент число исследуемых параметров можно сократить до 3 факторов.

Глава 2. Визуализация многомерных данных в SPSS

2.1. Методы для данных произвольной размерности

2.1.1. Матрица диаграмм рассеяния

Одним из основных методов визуализации в среде SPSS является матрица диаграмм рассеяния. На рисунке 3 приведен пример использования функции для факторов, полученных в первой главе.

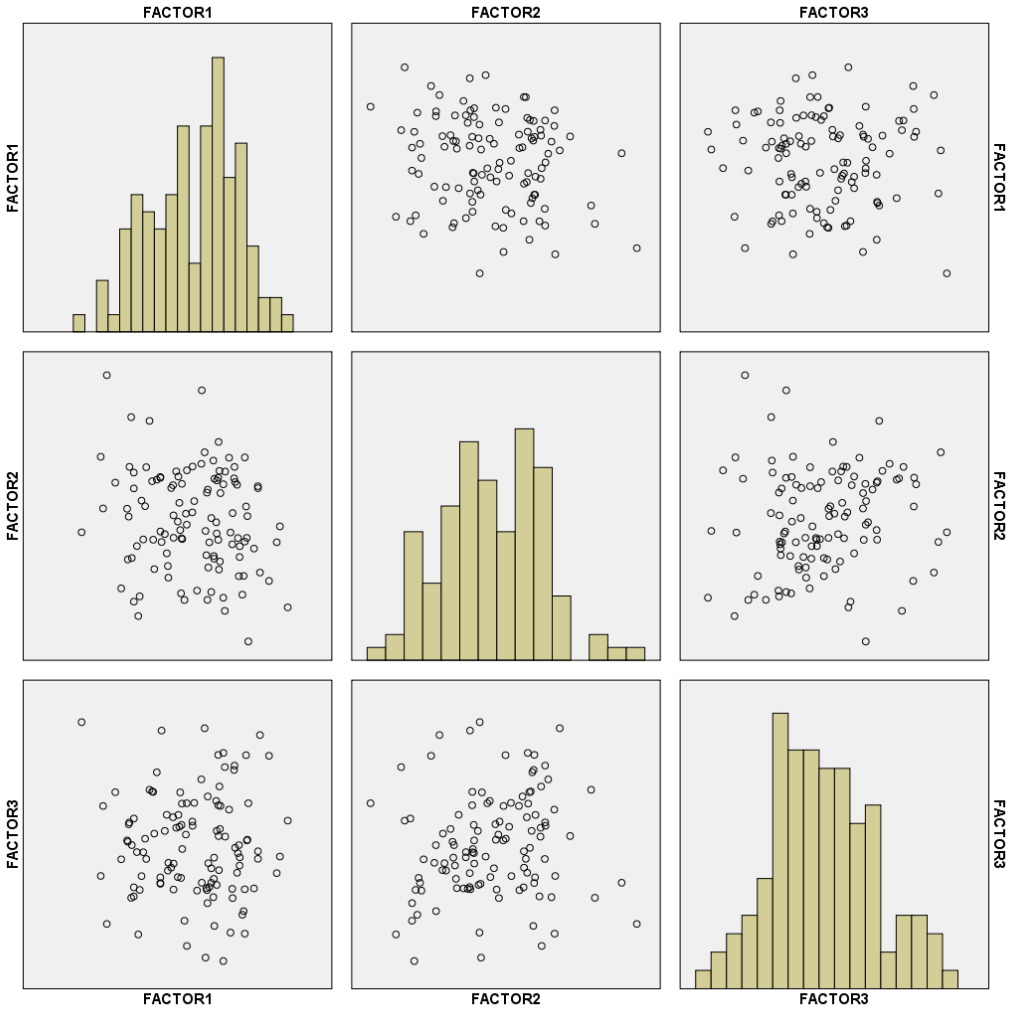


Рисунок 3. Матрица диаграмм рассеяния

На главной диагонали матрицы находятся гистограммы распределения переменной, а в качестве прочих элементов использованы диаграммы рассеяния точек, где оси Y соответствует переменная по строке, а оси X − переменная по столбцу.

График позволяет определить характер взаимосвязи между переменными (насколько сильно они коррелированы), а также дать предварительную оценку нормальности распределения параметров.

2.1.2. Перекрывающие диаграммы рассеяния

Для сопоставления множества диаграмм рассеяния используется метод перекрывающихся диаграмм рассеяния. В соответствие с ним, каждая диаграмма изображается в рамках одного и того же графика, но различным цветом. На рисунке 4 представление метода для рассмотренных ранее факторов.

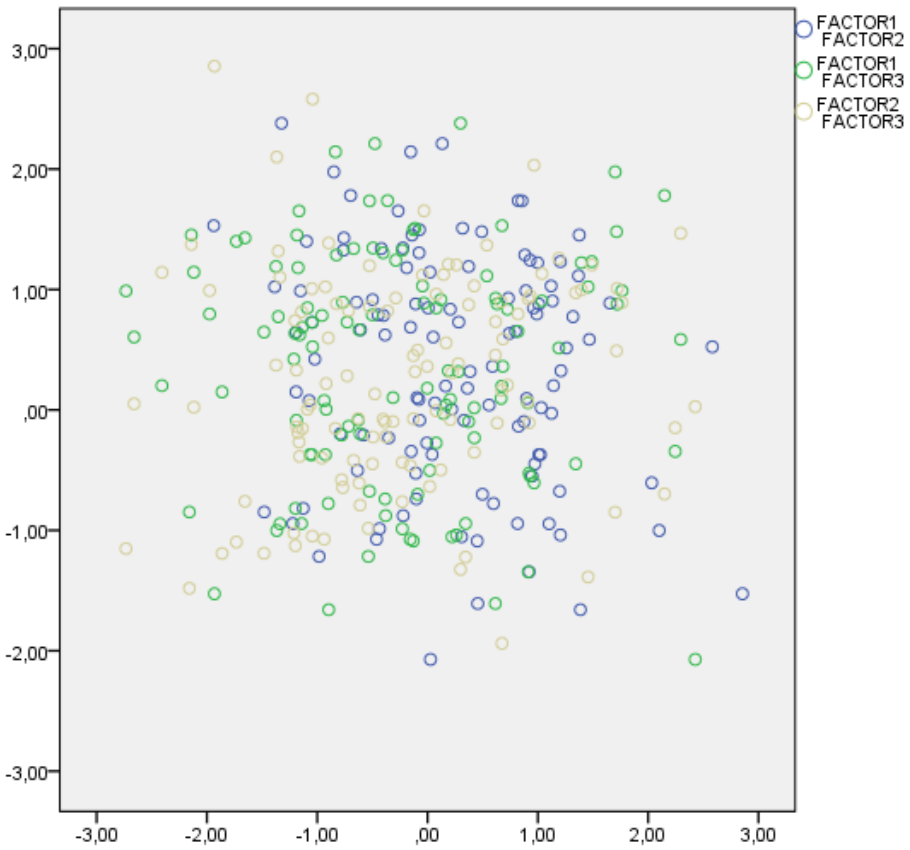


Рисунок 4. Перекрывающиеся диаграммы рассеяния

Данный метод позволяет определить различия и сходства во взаимосвязи различных переменных.

2.1.3. Метод параллельных координат

Суть метода параллельных координат состоит в представлении области значения переменных в виде вертикальных осей. На каждой из осей отмечается значение, соответствующее определенному наблюдению, а затем проводятся прямые, соединяющие точки, так, чтобы каждому наблюдению отвечал собственный цвет. Преимущество такого подхода в том, что можно легко сравнивать результаты различных наблюдений. Рисунок 5 иллюстрирует применение метода к исследуемым факторам.

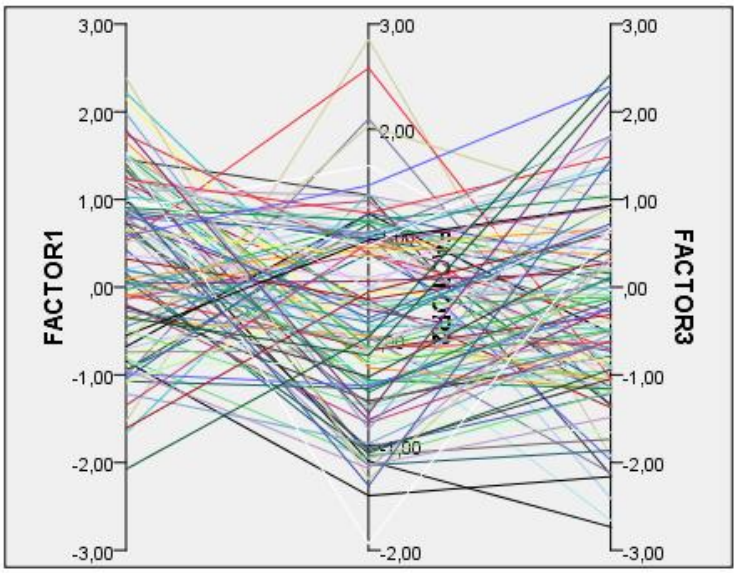


Рисунок 5. Метод параллельных координат

Данный пример хорошо иллюстрирует тот факт, что с увеличением количества наблюдений эффективность метода падает.

2.2. Методы для трехмерных данных

В результате исследования были получены три фактора. Таким образом, могут быть использованы трехмерные методы визуализации.

Однако эти методы не являются универсальными, так как ограничены тремя измерениями, поэтому в данной работе им не уделяется особого внимания, а приводится только их перечисление.

- Трехмерная диаграмма рассеяния;
- Поверхность;
- Диаграмма с пузырями.

2.3. Выводы из второй главы

В рамках пакета программ SPSS Statistics не реализован функционал, способный качественно справится с задачей визуализации многомерных данных [10], [11].

Из всех многомерных методов произвольной размерности только метод параллельных координат дает незначительное количество информации об исходных данных, а также перекрывающиеся диаграммы рассеяния позволяют сделать вывод о том, что у всех переменных имеется сходный характер взаимосвязи между собой.

Заключение

В ходе работы был выработан следующий алгоритм редукции многомерных данных.

1. Выделение параметров при помощи исследования корреляционных зависимостей переменных;
2. Построение многомерной регрессионной модели для прогнозирования значения исследуемого параметра;
3. Снижение размерности исходных данных при помощи метода главных компонент.

В работе рассмотрен пример применения алгоритма к базе данных Европейского Социального Исследования. В результате, был исследован вопрос об удовлетворенности граждан страны жизнью, построена регрессионная модель для предсказания значения «удовлетворенности», а также получены новые переменные, которые могут быть однозначно интерпретированы и использоваться вместо большего числа исходных параметров.

В заключении, можно сделать вывод о том, что использование сложных инструментов статистического анализа и дорогостоящего программного обеспечения оправдано лишь в том случае, когда важна высокая точность результата, оправдывающая все вычислительные затраты, или есть основания полагать, что существует некоторая сложная взаимосвязь между переменными. В противном же случае, разумно использовать более простые методы, предоставляющие достаточную степень точности.

Список литературы

1. Ресурсный центр «Вычислительный центр СПбГУ». http://www.cc.spbu.ru

2. About ESS. http://www.europeansocialsurvey.org/about/

3. Бююль А., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб.: ДиаСофтЮп, 2005. 608 с.

4. Орлова И. В., Концевая Н. А. и др. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS / под ред. И. В. Орловой. М.: Вузовский учебник, 2009. 320 с.

5. Сошникова Л. А., Тамашевич В. Н. и др. Многомерный статистический анализ в экономике: Учеб. пособие для вузов / под ред. проф. В. Н. Тамашевича. М.: ЮНИТИ-ДАНА, 1999. 598 с.

6. Буре В. М., Парилина Е. М. Теория вероятностей и математическая статистика. СПб.: Изд-во Лань, 2013. 416 с.

7. LearnSPSS: Порядковая регрессия.

http://www.learnspss.ru/hndbook/glava16/cont11.htm

8. LearnSPSS: Факторный анализ
http://www.learnspss.ru/hndbook/glava19/cont3.htm

9. Ким Дж., Мюллер Ч. и др. Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. 216 с.

10. Зиновьев А.Ю. Визуализация многомерных данных. Красноярск: Изд-во КГТУ, 2000. 168 с.

11. Бондарев А.Е., Галактионов В.А. Анализ многомерных данных в задачах многопараметрической оптимизации с применением методов визуализации // Научная визуализация, 2012. Т. 4, № 2. С. 1-13.